(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2008-257511
(P2008-257511A)
(43)公開日　平成20年10月23日(2008.10.23)

| (51)Int.Cl. | | F I | | テーマコード（参考） |
|---|---|---|---|---|
| G06F 17/28 | (2006.01) | G06F 17/28 | U | 5B075 |
| G06F 17/30 | (2006.01) | G06F 17/30 | 210A | 5B091 |

審査請求 未請求　請求項の数 10　ＯＬ　（全 24 頁）

(21)出願番号　特願2007-99403(P2007-99403)
(22)出願日　平成19年4月5日(2007.4.5)

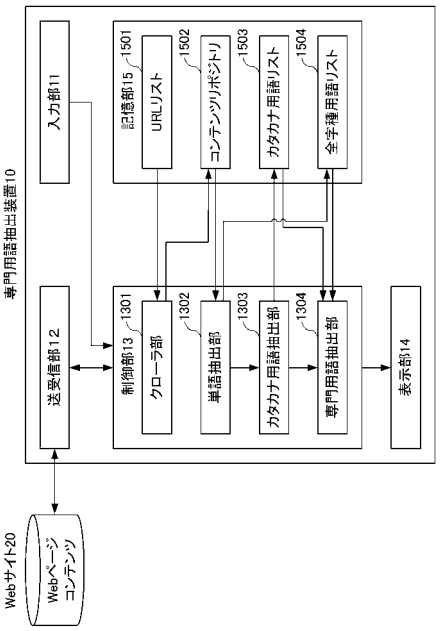
(71)出願人　500257300
ヤフー株式会社
東京都港区六本木六丁目１０番１号
(74)代理人　100106002
弁理士　正林　真之
(72)発明者　萩原　健
東京都港区六本木６丁目１０番１号　ヤフー株式会社内
(72)発明者　増山　毅司
東京都港区六本木６丁目１０番１号　ヤフー株式会社内
(72)発明者　本野　秀樹
東京都港区六本木６丁目１０番１号　ヤフー株式会社内

Ｆターム(参考)　5B075 NK32

最終頁に続く

(54)【発明の名称】専門用語抽出装置、方法及びプログラム

(57)【要約】
【課題】Ｗｅｂドキュメントから専門用語を自動的に抽出する専門用語抽出装置を提供すること。
【解決手段】本装置のクロール部が、専門分野ごとのＵＲＬリストを用いて、Ｗｅｂページのコンテンツを収集する。ＵＲＬリストは、常にＵｐ－Ｔｏ－Ｄａｔｅに更新する。次に、本装置の単語抽出部によって、収集されたＷｅｂページのコンテンツのテキストを形態素解析し品詞に分類して、カタカナ語彙と全字種の語彙を抽出する。この際、助詞や接続詞など専門用語になりにくい品詞は抽出対象から除外する。そして、本装置のカタカナ用語抽出部によって、抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて、重要度の計算を行い重要度の高いカタカナ用語を抽出する。さらに、専門用語抽出部によって、カタカナ用語と、先に抽出された全字種の語彙との共起ヒット情報を計算して、専門用語を抽出する。
【選択図】図２



【特許請求の範囲】
【請求項１】
Ｗｅｂページから専門用語を抽出する専門用語抽出装置であって、
専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するクローラ部と、
前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出部と、
前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出部と、
前記カタカナ用語抽出部によって抽出されたカタカナ用語と、前記単語抽出部によって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出部と、
を備えた専門用語抽出装置。
【請求項２】
前記カタカナ用語抽出部は、前記ＦＬＲ法に代えて、Ｃ－Ｖａｌｕｅ法を用いる、請求項１に記載の装置。
【請求項３】
前記カタカナ用語抽出部は、前記ＦＬＲ法に代えて、ＭＣ－Ｖａｌｕｅ法を用いる、請求項１に記載の装置。
【請求項４】
前記専門用語抽出部は、前記シンプソン係数値に代えて、相互情報量値を用いる、請求項１乃至３に記載の装置。
【請求項５】
前記専門用語抽出部は、前記シンプソン係数値に代えて、ダイス係数値を用いる、請求項１乃至３に記載の装置。
【請求項６】
前記専門用語抽出部は、前記シンプソン係数値に代えて、ジャガード係数値を用いる、請求項１乃至３に記載の装置。
【請求項７】
前記専門用語抽出部は、前記シンプソン係数値に代えて、コサイン類似度値を用いる、請求項１乃至３に記載の装置。
【請求項８】
前記専門用語として、アダルト専門分野における掲載禁止用語を抽出する請求項１乃至７に記載の装置。
【請求項９】
Ｗｅｂページから専門用語を抽出するための方法であって、
専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するステップと、
前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出ステップと、
前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出ステップと、
前記カタカナ用語抽出ステップによって抽出されたカタカナ用語と、前記単語抽出ステップによって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出ステップと、
を含む方法。
【請求項１０】
Ｗｅｂページから専門用語を抽出するためのコンピュータ・プログラムであって、
コンピュータに、
専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するステップと、

前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出ステップと、
前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出ステップと、
前記カタカナ用語抽出ステップによって抽出されたカタカナ用語と、前記単語抽出ステップによって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出ステップと、
を実行させるコンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ドキュメントからの専門用語抽出装置、方法及びプログラムに関する。特に、Ｗｅｂドキュメントからの専門用語抽出装置、方法及びプログラムに関する。
【背景技術】
【０００２】
様々な分野において、専門用語のデータベースを作成したり、データベースから専門用語を検索したりするために、専門分野のドキュメントから専門用語の抽出が行われている。従来、専門用語の抽出は当該分野の専門家が人手でドキュメントを精査し、抽出していたが、その作業を自動化するための試みが複数なされている。例えば、非特許文献１には、単名詞を含む単名詞バイグラムの左右に連接する単名詞を抽出し、その頻度を基にスコアリングを行い、専門用語を抽出する方法が開示されている。又、特許文献１には、大量の専門用語が抽出される分野において、専門用語辞書を最新状態にメンテナンスするために、ある用語の関連語の同族語、類似語の同族語を抽出することで、多様な周辺語彙を網羅的に情報収集し、新語登録などのメンテナンス作業を効率化する方法が開示されている。
【非特許文献１】出現頻度と連接頻度に基づく専門用語抽出、湯本他、自然言語処理、１０（１）２７－４５，２００３年１月
【特許文献１】特開２００５－２２２２６３号公報
【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかしながら、特許文献１及び非特許文献１に記載の技術では共に、専門用語を抽出する対象ドキュメントが既にデータベースに保存されており、かつ専門用語と関連する分野のドキュメントである（特許文献１であれば、医学・生物分野、非特許文献１であれば、情報処理分野）ことを前提としている。そのため、対象とするドキュメント数が限定されて、高精度で専門用語を抽出することができた。しかし、対象をＷｅｂサイト全体に広げた場合、Ｗｅｂドキュメントは分野ごとに分類されていないという問題があり、専門用語を抽出する前に、まず対象とするＷｅｂドキュメントをＷｅｂ上から収集する必要がある。又、Ｗｅｂサイトは次々に更新されるという特徴があり、さらに企業や官公庁だけでなく、個人の趣味・嗜好の基に作成されるものも多く存在するため、学術論文などに比べてノイズとなる情報がドキュメント中に多く含まれている可能性が高く、上記の技術とは別の視点が必要となる。
【０００４】
本発明は、上記課題に鑑み、Ｗｅｂドキュメントから専門用語を自動的に抽出する専門用語抽出装置を提供することを目的とする。
【課題を解決するための手段】
【０００５】
本発明では以下のような解決手段を提供する。
【０００６】
（１）　Ｗｅｂページから専門用語を抽出する専門用語抽出装置であって、

専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するクローラ部と、
前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出部と、
前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出部と、
前記カタカナ用語抽出部によって抽出されたカタカナ用語と、前記単語抽出部によって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出部と、
を備えた専門用語抽出装置。

【０００７】

（１）の構成によれば、まず、本装置に備えられたクロール部が、専門分野ごとに分けられたＵＲＬリストを用いて、Ｗｅｂページのコンテンツを収集（クロール）する。ＵＲＬリストは、常にＵｐ－Ｔｏ－Ｄａｔｅに更新する。次に、本装置の単語抽出部によって、収集されたＷｅｂページのコンテンツのテキストを形態素解析し品詞に分類して、カタカナ語彙と全字種の語彙を抽出する。この際、助詞や接続詞など専門用語になりにくい品詞は抽出対象から除外してよい。そして、本装置のカタカナ用語抽出部によって、カタカナ語彙から、ＦＬＲ法を用いて、重要度の計算を行い重要度の高いカタカナ用語を抽出する。さらに、専門用語抽出部によって、抽出されたカタカナ用語と、先に抽出された全字種の語彙との共起ヒット情報（２つの語彙が同じドキュメントで共起する度合い）を計算することによって専門用語を抽出する。なお、ＦＬＲ法とは、後述するように、連接頻度ＬＲ法（連接種類ＬＲ法）に、用語Ｗがコーパス（言語資料体）中に出現した頻度を加味したものである。

【０００８】

このように、まずカタカナ語彙に着目して重要度の高いカタカナ用語を求めるカタカナ用語抽出処理と、この重要度の高いカタカナ用語と全字種の語彙との共起ヒット情報による専門用語抽出処理を行うことによって、特にカタカナ語彙を含んだ専門用語（医薬分野、ＩＴ分野、ロボット工学分野、アダルト分野などの専門用語）に対して、膨大に存在するＷｅｂドキュメントから、Ｕｐ－ｔｏ－Ｄａｔｅに専門用語を自動的に抽出することが可能になる。

【０００９】

（２）　前記カタカナ用語抽出部は、前記ＦＬＲ法に代えて、Ｃ－Ｖａｌｕｅ法を用いる、（１）に記載の装置。Ｃ－Ｖａｌｕｅ法は、後述するように、用語Ｗを部分文字列として含むより長い用語の出現頻度を、用語Ｗを部分文字列として含むより長い用語の種類数で割った値を用語Ｗの出現頻度から補正した値を重要度とする方法である。

【００１０】

（２）の構成によれば、カタカナ用語抽出部において、ＦＬＲ法に代えて公知のＣ－Ｖａｌｕｅ法を用いることができる。

【００１１】

（３）　前記カタカナ用語抽出部は、前記ＦＬＲ法に代えて、ＭＣ－Ｖａｌｕｅ法を用いる、（１）に記載の装置。

【００１２】

（３）の構成によれば、カタカナ用語抽出部において、ＦＬＲ法に代えてＣ－Ｖａｌｕｅ法を改良したＭＣ－Ｖａｌｕｅ法（Ｍｏｄｉｆｉｅｄ　Ｃ－Ｖａｌｕｅ法）を用いることができる。

【００１３】

（４）　前記専門用語抽出部は、前記シンプソン係数値に代えて、相互情報量値を用いる、（１）乃至（３）に記載の装置。

【００１４】

（５）　前記専門用語抽出部は、前記シンプソン係数値に代えて、ダイス係数値を用いる、（１）乃至（３）に記載の装置。

【００１５】
（６）　前記専門用語抽出部は、前記シンプソン係数値に代えて、ジャガード係数値を用いる、（１）乃至（３）に記載の装置。
【００１６】
（７）　前記専門用語抽出部は、前記シンプソン係数値に代えて、コサイン類似度値を用いる、（１）乃至（３）に記載の装置。
【００１７】
（４）から（７）の構成によれば、専門用語抽出部において、ＴＦＩＤＦ法とシンプソン係数を組み合わせた方法以外にも共起ヒット情報を求める各種の手段（相互情報量値、ダイス係数値、ジャガード係数値、コサイン類似度値）を活用することができる。
【００１８】
（８）　前記専門用語として、アダルト専門分野における掲載禁止用語を抽出する（１）乃至（７）に記載の装置。
【００１９】
（８）の構成によれば、専門分野としてペアレンタルコントロールに着目し、有害サイト、特にアダルトサイトで使用されるような「掲載禁止用語」（以下、ＮＧ語彙とも呼ぶ）を抽出する。アダルトサイトは規制しても次々と新しいサイトが出現し、又ＮＧ語彙にはカタカナが多く使用されるので、このようなＮＧ語彙を含んだサイトのフィルタリングに本発明の手法が有効である。
【００２０】
（９）　Ｗｅｂページから専門用語を抽出するための方法であって、
　専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するステップと、
　前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出ステップと、
　前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出ステップと、
　前記カタカナ用語抽出ステップによって抽出されたカタカナ用語と、前記単語抽出ステップによって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出ステップと、
　を含む方法。
【００２１】
（９）の構成によれば、（１）と同様の作用効果を持つ発明を方法として提供できる。
【００２２】
（１０）　Ｗｅｂページから専門用語を抽出するためのコンピュータ・プログラムであって、
　コンピュータに、
　専門分野ごとに定められたＵＲＬリストに含まれたＵＲＬにアクセスし、前記Ｗｅｂページのコンテンツを収集するステップと、
　前記コンテンツを形態素解析し、カタカナ語彙と全字種語彙を抽出する単語抽出ステップと、
　前記抽出されたカタカナ語彙に対して、ＦＬＲ法を用いて前記カタカナ語彙からカタカナ用語を抽出するカタカナ用語抽出ステップと、
　前記カタカナ用語抽出ステップによって抽出されたカタカナ用語と、前記単語抽出ステップによって抽出された全字種語彙とを、ＴＦＩＤＦ値とシンプソン係数値を組み合わせた共起ヒット情報を計算して、専門用語を抽出する専門用語抽出ステップと、
　を実行させるコンピュータ・プログラム。
【００２３】
（１０）の構成によれば、（１）と同様の作用効果を持つ発明をコンピュータ・プログラムとして提供できる。

【発明の効果】
【００２４】
本発明によれば、カタカナ語彙が専門用語として多く使用される専門分野において、次々と更新されるＷｅｂサイト上の膨大なドキュメント群から、人手で精査することなく自動で専門用語抽出を行うことができる。
【発明を実施するための最良の形態】
【００２５】
以下、本発明の実施形態について図を参照しながら説明する。
【００２６】
［システムの全体構成］
図１は、本発明の一実施形態に係るシステム１の全体構成を示す図である。
【００２７】
本実施形態におけるシステム１は、テキストや画像などを含んだドキュメントデータ（例えば、インターネットやイントラネット上のＷｅｂページ）の解析を行い、ドキュメントデータに含まれる用語群を抽出して、該用語群から専門用語の抽出を行うシステムである。本システム１では、専門用語抽出装置１０が、通信ネットワーク３０を介して、様々なＷｅｂサイト２０と接続される。専門用語抽出装置１０は、専用装置であっても、他の目的のサーバ上に実現してもよい。なお、専門用語抽出装置１０のハードウェアの数に制限はなく、必要に応じて、１又は複数のハードウェアで構成してよい。
【００２８】
Ｗｅｂサイト２０は、Ｗｅｂページを蓄積しており、通信ネットワーク３０、例えば、インターネットなどのネットワークを通じて、これらの情報をユーザの端末に送信する機能を有している。なお、個人や会社のホームページなどのＷｅｂページ群、又はＷｅｂページ群が置いてあるインターネット、又はイントラネット上の場所を、Ｗｅｂサイトという。
【００２９】
通信ネットワーク３０は、例えば、インターネットであり、通信回線は有線により実現するものだけではなく、アクセスポイントを介して無線ＬＡＮにより実現するものなど、本発明の技術的思想に合致するものであれば様々な通信技術により実現される。
【００３０】
専門用語抽出装置１０は、専門分野ごとのＵＲＬリスト１５０１（ａ、ｂ、ｃ、ｄ、・・・）にあるＵＲＬのＷｅｂサイト２０を参照し、該Ｗｅｂサイト２０のＷｅｂページデータ（コンテンツ）を、通信ネットワーク３０を介して収集する。そして、収集したＷｅｂページデータをコンテンツリポジトリ１５０２に記憶する。さらに、収集したＷｅｂページに含まれるテキストデータを形態素解析して、語彙を抽出し、専門用語を抽出する機能を備える。
【００３１】
ここで、ＵＲＬリスト１５０１は、管理者が、特定の分野のＷｅｂサイト２０のＵＲＬをリストにすることによって与えられるものとする。例えば、特定の分野とは、情報処理分野のＷｅｂサイト２０（ＵＲＬリスト１５０１ａ）、医療・生物分野のＷｅｂサイト２０（ＵＲＬリスト１５０１ｂ）、アダルト専門分野のＷｅｂサイト２０（ＵＲＬリスト１５０１ｃ）、又はロボット工学関連分野のＷｅｂサイト２０（ＵＲＬリスト１５０１ｄ）などである。こうすることで、特定の分野における専門用語を抽出することができる。ここでは、ＵＲＬリスト１５０１が複数ある例を示しているが、１つのＵＲＬリスト１５０１に、ＵＲＬと特定の分野を関連付けて記憶することで実現してもよい。
【００３２】
なお、アダルト専門分野のＷｅｂサイト２０（ＵＲＬリスト１５０１ｃ）から専門用語を抽出するということは、公序良俗に反するような用語を抽出することである。そして、抽出した用語を掲載禁止用語（ＮＧワード）とし、このＮＧワードを含むＷｅｂサイトの検索に用いたり、有害サイトの特定に用いることができる。

【００３３】
［専門用語抽出装置の機能ブロック］
図２は、本発明の一実施形態に係る専門用語抽出装置１０の機能ブロック図である。
【００３４】
専門用語抽出装置１０は、主として入力部１１、送受信部１２、制御部１３、表示部１４、及び記憶部１５により構成される。入力部１１は、キーボード及びマウスなどの入力装置を含み、専門用語抽出装置１０に対する管理者などからの入力を受け付ける機能を有している。又、送受信部１２は、任意の通信インターフェイスを含み、装置からリクエストをＷｅｂサイト２０に送信する機能、及びＷｅｂサイト２０のＷｅｂページデータを受信する機能を有している。さらに、制御部１３は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）を含み、専門用語抽出装置１０を制御する機能を有している。そして、表示部１４は、ブラウン管表示装置（ＣＲＴ）や液晶ディスプレイ（ＬＣＤ）などの表示装置を含み、データを表示する機能を有している。又さらに、記憶部１５は、ハードディスクなどの内部又は外部の記憶装置を含み、データを記憶する機能を有している。
【００３５】
専門用語抽出装置１０の制御部１３は、クローラ部１３０１、単語抽出部１３０２、カタカナ用語抽出部１３０３、及び専門用語抽出部１３０４を有している。クローラ部１３０１は、通信ネットワーク３０を介して、Ｗｅｂページなどのドキュメントデータを収集する。なお、クローラとは一般的に検索ロボットともいわれ、通信ネットワーク３０を通じて、Ｗｅｂサイト２０からＷｅｂページデータを収集するプログラムである。そして、クローラが、Ｗｅｂサイトを探し出す手段や、対象とするＷｅｂページデータの種類は様々であり、クローラの管理者の設定により、収集されるＷｅｂページデータの種類や分野も異なる。
【００３６】
又、単語抽出部１３０２は、ドキュメント中のテキストを形態素解析して、単語を抽出し、カタカナ語彙と、全字種の語彙とに分けて、全字種の語彙を全字種用語リスト１５０４に記憶する。そして、カタカナ用語抽出部１３０３は、カタカナ語彙の用語ごとに重要度（後述）を計算し、管理者の設定する閾値以上の用語を抽出し、カタカナ用語リスト１５０３に記憶する。さらに、専門用語抽出部１３０４は、カタカナ用語リスト１５０３と、全字種用語リスト１５０４とにおいて共起の強い用語を専門用語として抽出する。
【００３７】
専門用語抽出装置１０の記憶部１５は、ＵＲＬリスト１５０１、コンテンツリポジトリ１５０２、カタカナ用語リスト１５０３、及び全字種用語リスト１５０４を含んで構成される。ＵＲＬリスト１５０１は、クローラ部１３０１によるＷｅｂページデータ収集先のＷｅｂサイト２０のＵＲＬを記憶する。又、コンテンツリポジトリ１５０２は、クローラ部１３０１により収集されたＷｅｂページデータを記憶する。そして、カタカナ用語リスト１５０３は、カタカナ用語を記憶する。さらに、全字種用語リスト１５０４は、全字種の語彙を記憶する。
【００３８】
［専門用語抽出処理］
図３は、本発明の一実施形態に係る専門用語抽出処理のフローチャートである。
【００３９】
まず、ステップＳ１０１では、専門用語抽出装置１０の制御部１３が、送受信部１２を介して、クローラ部１３０１により、Ｗｅｂページなどのドキュメントデータを収集する。なお、記憶部１５のＵＲＬリスト１５０１に含まれたＵＲＬに対する、Ｗｅｂサイト２０のＷｅｂページデータを収集してもよい。
【００４０】
次に、ステップＳ１０２では、クローラ部１３０１が、収集したドキュメントデータを、コンテンツリポジトリ１５０２に記憶する。
【００４１】

次に、ステップＳ１０３では、単語抽出部１３０２が、コンテンツリポジトリ１５０２から、ドキュメントデータを読み込む。

【００４２】

次に、ステップＳ１０４では、単語抽出部１３０２が、ドキュメントデータのテキストを形態素解析する。ここで、形態素解析とは、文を形態素（例えば、言語で意味を持つ最小単位）の列に分割し、接続詞や助詞を取り除く。形態素解析には様々な公知の手法があるが、いずれの手法を用いてもよい。

【００４３】

次に、ステップＳ１０５では、単語抽出部１３０２が、ドキュメントデータのテキストを形態素解析した結果の中から、全字種の語彙を抽出する。そして、全字種の語彙を、全字種用語として、記憶部１５の全字種用語リスト１５０４に記憶する。

【００４４】

次に、ステップＳ１０６では、上述のステップＳ１０５を行うと共に、単語抽出部１３０２が、ドキュメントデータのテキストを形態素解析した結果の中から、カタカナ語彙を抽出する。

【００４５】

次に、ステップＳ１０７では、カタカナ用語抽出部１３０３が、カタカナ語彙の用語ごとに重要度（後述）を計算し、管理者の設定する閾値以上の用語を特定する。なお、カタカナ用語特定処理の詳細については、図４で後述する。

【００４６】

次に、ステップＳ１０８では、カタカナ用語抽出部１３０３が、カタカナ語彙に対してカタカナ用語特定処理を行い特定した用語群を抽出して、記憶部１５のカタカナ用語リスト１５０３に記憶する。

【００４７】

次に、ステップＳ１０９では、専門用語抽出部１３０４が、カタカナ用語リスト１５０３を用いて、全字種用語リスト１５０４の用語群の中から専門用語を特定する。なお、専門用語特定処理の詳細については、図５で後述する。

【００４８】

次に、ステップＳ１１０では、専門用語抽出部１３０４が、全字種用語リスト１５０４から専門用語特定処理を行い特定した用語群を、専門用語として抽出する。そして、抽出した専門用語と共に、カタカナ用語リスト１５０３の用語を専門用語として、専門用語辞書に登録してもよい。

【００４９】

図４は、本発明の一実施形態に係るカタカナ用語特定処理のフローチャートである。

【００５０】

まず、ステップＳ１７１では、専門用語抽出装置１０の制御部１３が、カタカナ用語抽出部１３０３により、カタカナ語彙について用語ごとに重要度を計算する。なお、重要度の計算方法は、ＦＬＲ（Ｆｒｅｑｕｅｎｃｙ　Ｌｅｆｔ　Ｒｉｇｈｔ）法、Ｃ－Ｖａｌｕｅ（Ｃｏｌｌｏｃａｔｉｏｎ－Ｖａｌｕｅ）法、ＭＣ－Ｖａｌｕｅ（Ｍｏｄｉｆｉｅｄ　Ｃｏｌｌｏｃａｔｉｏｎ－Ｖａｌｕｅ）法などがあるので以下説明する。

【００５１】

ＦＬＲ法は、連接頻度ＬＲ法又は連接種類ＬＲ法に、用語Ｗがドキュメントデータ中に出現した頻度Ｆを加味する方法である。詳細は（非特許文献１）を参照。連接頻度ＬＲ法は、語彙を走査し、用語Ｗを構成する単語について、該単語の左右それぞれに単語が出現する回数を計算する。又、連接種類ＬＲ法は、単語の左右それぞれに何種類の単語が出現するかをカウントする。ここで、例えば、カタカナ語彙中の用語「サーバシステム、コンピュータシステム、オープンシステム」があり、構成する単語を分けると（サーバ｜システム）、（コンピュータ｜システム）、（オープン｜システム）となり、単語「システム」の左に単語が３回出現したので、単語「システム」の連接頻度ＬＲ法での左方スコアはＬ（システム）＝３となる。又、単語「システム」の左に単語が３種類出現したので、連

接種類ＬＲ法での左方スコアはＬ（システム）＝３となる。

【００５２】

一般に、単語ｗ１、ｗ２、・・・ｗｎが連なって構成する用語Ｗ＝ｗ１、ｗ２、・・・ｗｎについて、連接頻度ＬＲ法又は連接種類ＬＲ法の用語Ｗのスコア ＬＲ（Ｗ）が、数１のように定義される。

【数１】

$$LR(W) = (\prod_{i=1}^{n} (L(w_i)+1)(R(w_i)+1))^{\frac{1}{2n}}$$

ｎ：単語数

Ｌ（Ｗｉ）、Ｒ（Ｗｉ）：単語Ｗｉの左右それぞれに単語が出現する回数又は種類数

【００５３】

そして、連接頻度ＬＲ（Ｗ）又は連接種類ＬＲ（Ｗ）に、用語Ｗがドキュメントデータ中に出現した頻度Ｆ（Ｗ）を加味した、重要度ＦＬＲ（Ｗ）が、数２のように定義される。

【数２】

$$FLR(W) = F(W) \times LR(W)$$

Ｆ（Ｗ）：用語Ｗのドキュメントデータ中の出現頻度

ＬＲ（Ｗ）：用語Ｗの連接頻度ＬＲ又は連接種類ＬＲ

【００５４】

又、Ｃ－Ｖａｌｕｅ法は、用語Ｗ＝ｗ１、ｗ２、・・・ｗｎについて、重要度Ｃ－Ｖａｌｕｅ（Ｗ）が、数３のように定義される。Ｃ－Ｖａｌｕｅ法についての詳細は（Ｋａｔｅｒｉｎａ Ｔ．Ｆｒａｎｔｚｉ ａｎｄ Ｓｏｐｈｉａ Ａｎａｎｉａｄｏｕ．Ｅｘｔｒａｃｔｉｎｇ ｎｅｓｔｅｄ ｃｏｌｌｏｃａｔｉｏｎｓ．Ｉｎ ＣＯＬＩＮＧ‘９６，ｐｐ．４１－４６，１９９６．）を参照。

【数３】

$$C-Value(W) = (n-1) \times (F(W) - \frac{T(W)}{C(W)})$$

ｎ：単語数

Ｔ（Ｗ）：用語Ｗを部分文字列として含むより長い用語の出現頻度

Ｃ（Ｗ）：用語Ｗを部分文字列として含むより長い用語の種類数

Ｆ（Ｗ）：用語Ｗのドキュメントデータ中の出現頻度

【００５５】

なお、Ｃ－Ｖａｌｕｅ法は、ｎ＝１のとき（用語が単一の単語だけからなるとき）０（ゼロ）になり、適切な重要度を示さない。そこで、ＭＣ－Ｖａｌｕｅ法では、ｎ＝１の場合でも重要度を計算できるよう、（ｎ－１）の代わりにｎを用いている。ここで、用語Ｗ＝ｗ１、ｗ２、・・・ｗｎについて、重要度ＭＣ－Ｖａｌｕｅ（Ｗ）が、数４のように定義される。ＭＣ－Ｖａｌｕｅ法についての詳細は（非特許文献１）を参照。

【数４】

$$MC-Value(W)=n\times(F(W)-\frac{T(W)}{C(W)})$$

ｎ：単語数
Ｔ（Ｗ）：用語Ｗを部分文字列として含むより長い用語の出現頻度
Ｃ（Ｗ）：用語Ｗを部分文字列として含むより長い用語の種類数
Ｆ（Ｗ）：用語Ｗのドキュメントデータ中の出現頻度

【００５６】
次に、ステップＳ１７２では、カタカナ用語抽出部１３０３が、カタカナ語彙から、管理者が設定した閾値以上の重要度の用語を特定する。このようにして、カタカナ語彙から、カタカナの専門用語を特定することができる。

【００５７】
図５は、本発明の一実施形態に係る専門用語特定処理のフローチャートである。

【００５８】
まず、ステップＳ１９１では、専門用語抽出装置１０の制御部１３が、専門用語抽出部１３０４により、カタカナ用語リスト１５０３を用いて、全字種用語リスト１５０４のそれぞれの用語について、共起ヒット情報を計算する。ここで、共起ヒット情報の計算方法は、シンプソン係数値とＴＦ・ＩＤＦ（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ・Ｉｎｖｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）法とを用いる。

【００５９】
なお、シンプソン係数値は、用語と用語の共起の強さを測る尺度であり、スコアが０～１の範囲で、高いほど共起が強い。そして、カタカナ用語Ｘと全字種用語Ｙについての、シンプソン係数値Ｒ（Ｘ，Ｙ）が、数５のように定義される。

【数５】

$$R(X,Y)=\frac{|X\cap Y|}{\min(|X|,|Y|)}$$

｜Ｘ｜：カタカナ用語Ｘの単独ヒット数
｜Ｙ｜：全字種用語Ｙの単独ヒット数
｜Ｘ∩Ｙ｜：カタカナ用語Ｘと全字種用語ＹのＡＮＤ検索でのヒット数

【００６０】
次に、ステップＳ１９２では、専門用語抽出部１３０４が、共起ヒット情報を基に、管理者が設定した閾値以上の共起の強さを持つ用語を、専門用語として特定する。

【００６１】
［実施例１］
以下、専門分野として「アダルト専門分野」を例に、カタカナ用語抽出部１３０３による、重要度ＦＬＲの計算方法を具体的に説明する。図６は、実施例１に係るアダルト専門分野のドキュメントデータのカタカナ語彙を示す図である。図６（ａ）は、カタカナ語彙中の単語「セックス」を含む用語群である。図６（ｂ）は、単語「セックス」の左右連接単語の出現頻度である。図６（ｃ）は、カタカナ語彙中の単語「パートナー」を含む用語群である。図６（ｄ）は、単語「パートナー」の左右連接単語の出現頻度である。ここで、単語「セックス」と単語「パートナー」についてＦＬＲ法による重要度を計算する。

【００６２】
まず、連接頻度法に基づく、重要度ＦＬＲを計算する。カタカナ語彙中の単語「セックス」を含む用語群（図６（ａ））において、単語「セックス」の出現頻度Ｆ（セックス）

はｎ＝３である。そして、図６（ｂ）に示すように、単語「セックス」の左連接単語は、「アナルセックス（３）、テレホンセックス（１）、オーラルセックス（１）」であることから、頻度Ｌ（セックス）がｉ＝５となる。又、右連接単語は、「セックスパートナー（２）、セックスレス（１）」であることから、頻度Ｒ（セックス）がｉ＝３となる。ここで、連接頻度に基づく重要度ＦＬＲ（セックス）を計算する。

【数６】

$$
\begin{aligned}
FLR(\text{セックス}) &= F(\text{セックス}) \times LR(\text{セックス}) \\
&= F(\text{セックス}) \times \prod_{i=1}^{1} (L(\text{セックス})+1)(R(\text{セックス})+1)^{\frac{1}{2}} \\
&= F(\text{セックス}) \times \sqrt{(L(\text{セックス})+1)(R(\text{セックス})+1)} \\
&= 3 \times \sqrt{(5+1)(3+1)} \\
&= 3 \times \sqrt{24} \\
&\approx 14.70
\end{aligned}
$$

このようにして、連接頻度法に基づく重要度ＦＬＲ（セックス）は１４．７０と計算される。

【００６３】

続いて、カタカナ語彙中の単語「パートナー」を含む用語群（図６（ｃ））において、単語「パートナー」の出現頻度Ｆ（パートナー）はｎ＝２である。そして、図６（ｄ）に示すように、単語「パートナー」の左連接単語は、「セックスパートナー（２）」であることから、頻度Ｌ（セックス）がｉ＝２となる。又、右連接単語は、「パートナーリレーション（１）」であることから、頻度Ｒ（セックス）がｉ＝１となる。ここで、連接頻度に基づく重要度ＦＬＲ（パートナー）を計算する。

【数７】

$$
\begin{aligned}
FLR(\text{パートナー}) &= F(\text{パートナー}) \times LR(\text{パートナー}) \\
&= F(\text{パートナー}) \times \prod_{i=1}^{1} (L(\text{パートナー})+1)(R(\text{パートナー})+1)^{\frac{1}{2}} \\
&= F(\text{パートナー}) \times \sqrt{(L(\text{パートナー})+1)(R(\text{パートナー})+1)} \\
&= 2 \times \sqrt{(2+1)(1+1)} \\
&= 2 \times \sqrt{6} \\
&\approx 4.9
\end{aligned}
$$

このようにして、連接頻度法に基づく重要度ＦＬＲ（パートナー）は４．９と計算される。

【００６４】

次は、連接種類法に基づく、重要度ＦＬＲを計算する。カタカナ語彙中の単語「セックス」を含む用語群（図６（ａ））において、単語「セックス」の出現頻度Ｆ（セックス）＝３である。そして、図６（ｂ）に示すように、単語「セックス」の左連接単語は、「アナルセックス、テレホンセックス、オーラルセックス」であることから、種類Ｌ（セックス）がｉ＝３となる。又、右連接単語は、「セックスパートナー、セックスレス」であることから、種類Ｒ（セックス）がｉ＝２となる。ここで、連接種類に基づく重要度ＦＬＲ（セックス）を計算する。

【数８】

$$
\begin{aligned}
FLR(\text{セックス}) &= F(\text{セックス}) \times LR(\text{セックス}) \\
&= F(\text{セックス}) \times \prod_{i=1}^{1} (L(\text{セックス})+1)(R(\text{セックス})+1)^{\frac{1}{2}} \\
&= F(\text{セックス}) \times \sqrt{(L(\text{セックス})+1)(R(\text{セックス})+1)} \\
&= 3 \times \sqrt{(3+1)(2+1)} \\
&= 3 \times \sqrt{12} \\
&\approx 10.4
\end{aligned}
$$

このようにして、連接種類法に基づく重要度ＦＬＲ（セックス）は１０．４と計算される。

【００６５】

続いて、カタカナ語彙中の単語「パートナー」を含む用語群（図６（ｃ））において、単語「パートナー」の出現頻度Ｆ（パートナー）＝２である。そして、図６（ｄ）に示すように、単語「パートナー」の左連接単語は、「セックスパートナー」であることから、種類Ｌ（セックス）がｉ＝１となる。又、右連接単語は、「パートナーリレーション」であることから、種類Ｒ（セックス）がｉ＝１となる。ここで、連接種類法に基づく重要度ＦＬＲ（パートナー）を計算する。

【数９】

$$
\begin{aligned}
FLR(\text{パートナー}) &= F(\text{パートナー}) \times LR(\text{パートナー}) \\
&= F(\text{パートナー}) \times \prod_{i=1}^{1} (L(\text{パートナー})+1)(R(\text{パートナー})+1)^{\frac{1}{2}} \\
&= F(\text{パートナー}) \times \sqrt{(L(\text{パートナー})+1)(R(\text{パートナー})+1)} \\
&= 2 \times \sqrt{(1+1)(1+1)} \\
&= 3 \times \sqrt{4} \\
&= 6
\end{aligned}
$$

このようにして、連接種類法に基づく重要度ＦＬＲ（パートナー）は６と計算される。

【００６６】

このように、ＦＬＲ法に基づき、重要度を計算することができる。そして、閾値以上の重要度の用語を、専門用語として特定する。ここで、例えば、連接頻度において、ＦＬＲ（セックス）が１４．７０、ＦＬＲ（パートナー）が４．９の場合、閾値を８に設定することで、単語「セックス」のみが専門用語として特定できる。又、連接種類において、ＦＬＲ（セックス）が１０．４、ＦＬＲ（パートナー）が６の場合、閾値を８に設定することで、単語「セックス」のみが専門用語として特定できる。こうすることにより、カタカナ語彙中の用語から、閾値以上の重要度の用語を、アダルト専門分野のカタカナの専門用語として特定できる。

【００６７】

次に、専門用語抽出部１３０４による、共起ヒットの計算方法を具体的に説明する。図７は、実施例１に係る共起ヒットの具体例を示す図である。

【００６８】

まず、カタカナ用語リスト１５０３のカタカナ用語「セックス」と、全字種用語リスト１５０４の全字種用語「胸チラ」とについて、シンプソン係数値を計算する。ここで、図７に示す、ドキュメントデータにおける、カタカナ用語「セックス」の単独ヒット数（検索して抽出された数）は７００９、全字種用語「胸チラ」の単独ヒット数は４５２、カタカナ用語「セックス」と全字種用語「胸チラ」とでＡＮＤ検索したヒット数は４１４であ

る。ここで、シンプソン係数値Ｒ（セックス，胸チラ）が、数１０のように計算される。
【数１０】

$$R(\text{セックス，胸チラ}) = \frac{|\text{セックス} \cap \text{胸チラ}|}{min(|\text{セックス}|, |\text{胸 チラ}|)} = 414/452 = 0.915$$
$$|\text{セックス}| = 7009$$
$$|\text{胸チラ}| = 452$$
$$|\text{セックス} \cap \text{胸チラ}| = 414$$

このことにより、カタカナ用語「セックス」と全字種用語「胸チラ」との共起の強さが０．９１５となり、１に近いので共起が強いことがわかる。
【００６９】
次に、カタカナ用語リスト１５０３のカタカナ用語「セックス」と、全字種用語リスト１５０４の全字種用語「週末」とについて、シンプソン係数値を計算する。ここで、図７に示す、ドキュメントデータにおける、カタカナ用語「セックス」の単独ヒット数は７００９、全字種用語「週末」の単独ヒット数は１０６３、カタカナ用語「セックス」と全字種用語「週末」とでＡＮＤ検索したヒット数は２７８である。ここで、シンプソン係数値Ｒ（セックス，胸チラ）は数１１のように計算される。
【数１１】

$$R(\text{セックス，週末}) = \frac{|\text{セックス} \cap \text{週末}|}{min(|\text{セックス}|, |\text{週 末}|)} = 278/1063 = 0.262$$
$$|\text{セックス}| = 7009$$
$$|\text{週末}| = 1063$$
$$|\text{セックス} \cap \text{週末}| = 278$$

このことにより、カタカナ用語「セックス」と全字種用語「週末」との共起の強さが０．２６２となり、０（ゼロ）に近いので共起が弱いことがわかる。
【００７０】
このようにして、カタカナ用語リスト１５０３のカタカナ用語と、全字種用語リスト１５０４の全字種用語とについて、シンプソン係数値を計算する。そして、全字種用語リスト１５０４の全字種用語を、シンプソン係数値で降順にソートし、専門用語を抽出するが、いくつかの問題点がある。ここで、シンプソン係数値の問題点と解決方法とについて、図８に基づき説明する。
【００７１】
図８は、実施例１に係る全字種用語リスト１５０４の全字種用語をシンプソン係数値で降順にソートした図である。はじめの行には全字種用語「風俗店」がシンプソン係数値＝１．０００であることが示されている。同様に、シンプソン係数値の降順に全字種用語が並ぶ。
【００７２】
ここで、シンプソン係数値の問題点の１つには、低頻度な全字種用語について、共起が少なくてノイズである場合が多いが、シンプソン係数値が高くなりやすい問題がある。例えば、図８の６行目　全字種用語「具体案」の０．６６７は、カタカナ用語リスト１５０３のカタカナ用語「セックス」との、シンプソン係数値を示す。ここでは、ドキュメントデータにおける、カタカナ用語「セックス」の単独ヒット数は７００９、全字種用語「具体案」の単独ヒット数は３、カタカナ用語「セックス」と全字種用語「具体案」とでＡＮＤ検索したヒット数は２である。ここで、シンプソン係数値は数１２のように計算されて

いる。

【数１２】

$$R(\text{セックス},\ \text{具体例}) = \frac{|\text{セックス} \cap \text{具体例}|}{min(|\text{セックス}|,\ |\text{具 体例}|)} = 2/3 = 0.667$$

$$|\text{セックス}| = 7009$$

$$|\text{具体案}| = 3$$

$$|\text{セックス} \cap \text{具体案}| = 2$$

このことにより、カタカナ用語「セックス」と全字種用語「具体案」との共起の強さが０．６６７となる。しかし、この場合、カタカナ用語「セックス」の単独ヒット数が７００９であるのに対して、全字種用語「具体案」の単独ヒット数が３と低頻度である。よって、共起が強いとはいえない。そこで、全字種用語の単独ヒット数について閾値を設けることで解決することができる。例えば、閾値を４に設定することにより、全字種用語「具体案」について全字種用語から除くことができる。

【００７３】

しかし、閾値を設定することにより、どのカタカナ用語とも共起するような全字種用語（いわゆる一般語）は、単独ヒット数が多く、シンプソン係数値が高くなりやすい問題がある。例えば、図８の１０行目　全字種用語「フリーウェア」の０．６１３は、カタカナ用語リスト１５０３のカタカナ用語「セックス」との、シンプソン係数値を示す。ここでは、ドキュメントデータにおける、カタカナ用語「セックス」の単独ヒット数は７００９、全字種用語「フリーウェア」の単独ヒット数は６２、カタカナ用語「セックス」と全字種用語「フリーウェア」とでＡＮＤ検索した単独ヒット数は３８である。ここで、シンプソン係数値は数１３のように計算されている。

【数１３】

$$R(\text{セックス},\ \text{フリーウェア}) = \frac{|\text{セックス} \cap \text{フリーウェア}|}{min(|\text{セックス}|,\ |\text{フリーウェア}|)} = 38/62 = 0.613$$

$$|\text{セックス}| = 7009$$

$$|\text{フリーウェア}| = 62$$

$$|\text{セックス} \cap \text{フリーウェア}| = 38$$

このことにより、カタカナ用語「セックス」と全字種用語「フリーウェア」との共起の強さが０．６１３となる。しかし、全字種用語「フリーウェア」は一般語であるので、全字種用語から除く。そこで、閾値を６３に設定することにより、全字種用語「フリーウェア」について全字種用語から除くことができるが、他の全字種用語も除かれてしまう。そこで、公知のＴＦ・ＩＤＦ法を用いて解決をする。

【００７４】

ＴＦ・ＩＤＦ法は、ドキュメントの特徴を示す単語を抽出する方法であり、ドキュメントデータの特定のページに偏って多く出現する単語ほど高スコアとなる。なお、単語Ｘについてのｔｆ・ｉｄｆ値は、数１４のように定義される。

【数１４】

$$TF \cdot IDF = TF \cdot \log\left(\frac{N}{DF}\right)$$

ＴＦ：単語Ｘの全ページ中の出現頻度
ＤＦ：単語Ｘのページ頻度（いくつのページに跨って出現したか）
Ｎ：総ページ数
ここで、具体的な例を示す。

【００７５】
まず全字種用語「胸チラ」が、ＴＦ値＝１４２３、ＩＤＦ値＝６．０５９である場合、ＴＦ・ＩＤＦ値は８６２２．９５３となる。又、全字種用語「フリーウェア」が、ＴＦ値＝９７、ＩＤＦ値＝７．７９９である場合、ＴＦ・ＩＤＦ値は７５６．５４２となる。ここで、全字種用語「胸チラ」はＴＦ・ＩＤＦ値が高くドキュメントデータの特定のページに偏って多く出現していることが分かる。そして、全字種用語「フリーウェア」はＴＦ・ＩＤＦ値が低いのでドキュメントデータ全体に、一般語として使われていることが分かる。このことにより、例えば全字種用語「フリーウェア」のような、どのカタカナ用語とも共起するような全字種用語について、ＴＦ・ＩＤＦ値を用い、閾値を設けることで全字種用語から除くことができる。ここで、専門用語特定処理にシンプソン係数値とＴＦ・ＩＤＦ法とを用いた具体的な例について、図９に基づき説明する。

【００７６】
図９は、実施例１に係る全字種用語リスト１５０４から抽出された全字種用語を示す図である。ここで、全字種用語は、第１キー：シンプソン係数値、第２キー：ＴＦ・ＩＤＦ値でソートしている。そして、シンプソン係数値を求める際に、全字種用語の単独ヒット数の閾値を５６に設定し、低頻度な全字種用語を除いてある。又、全字種用語のＴＦ・ＩＤＦ値について、閾値を７６０に設定し、どのカタカナ用語とも共起するような全字種用語を除いてある。このようにして、共起ヒット情報に基づいた、全字種用語を特定することができる。

【００７７】
ここで、専門用語特定処理後の全字種用語の具体的な例について、図１０に基づき説明する。

【００７８】
図１０は、実施例１に係る専門用語特定処理後の全字種用語を示す図である。

【００７９】
図１０に示すように、全字種用語リスト１５０４の全字種用語から、専門用語として、「風俗店、女王、風俗嬢、・・・」といった全字種用語が抽出されている。又、全字種用語であったが、単独ヒット数の閾値を設定することで、「具体案、介護士、やすみ、・・・」といった全字種用語を、専門用語とすることが回避されている。さらに、ＴＦ・ＩＤＦ値を用いて、「フリーウェア、行楽地、株投資、・・・」といった全字種用語を、専門用語とすることが回避されている。すなわち、抽出された専門用語は、アダルト専門分野のＷｅｂサイト２０に使用されている専門用語であるため、掲載禁止用語として用いることができる。そして、抽出した掲載禁止用語と共に、カタカナ用語リスト１５０３の用語を掲載禁止用語として、掲載禁止用語辞書に登録してもよい。

【００８０】
なお、本発明の専門用語抽出には、形態素解析を用いるが、形態素解析後の品詞の並びを参照して、連続した単語を抽出してもよい。つまり、単独では専門用語とならない単語でも、単語同士を組み合わせた場合に、掲載禁止用語となる専門用語を抽出する。例えば、単語「女子高生」と単語「画像」は、それぞれ一般的な用語であるが、２つの単語を組み合わせた用語「女子高生画像」を、全字種用語として抽出する。そして、アダルト専門分野のＷｅｂサイト２０のドキュメントデータにおいて、カタカナ用語「○○○」などとの共起の強さを計算し、専門用語として特定する。そして、全字種用語「女子高生画像」を掲載禁止用語として抽出できる。

【００８１】
［実施例２］
以下、専門分野として、ロボット工学関連分野を対象とした実施例を説明する。

【００８２】
専門用語抽出装置１０を含むシステム１の構成及び機能ブロックは、図２と同様である。ここでは、ＵＲＬリスト１５０１は、ロボット工学関連分野のＵＲＬが設定されたＵＲＬリスト１５０１ｄを使用する。
【００８３】
又、専門用語抽出処理における実施形態は、図３と同様である。ここでは、クローラ部１３０１が、ロボット工学関連分野のＵＲＬリスト１５０１ｄに基づき、ロボット工学関連分野のＷｅｂサイト２０のドキュメントデータを収集し、コンテンツリポジトリ１５０２に記憶する。そして、単語抽出部１３０２が、コンテンツリポジトリ１５０２のドキュメントデータを形態素解析し、カタカナ語彙と全字種の語彙を抽出する。ここで、全字種の語彙は、全字種用語リスト１５０４に記憶する。
【００８４】
次に、カタカナ用語抽出部１３０３が、カタカナ語彙の用語ごとに重要度を計算し、管理者の設定する閾値以上の用語を特定するカタカナ用語特定処理については、図４と同様である。そして、カタカナ語彙において特定した用語をカタカナ用語リスト１５０３に記憶する。
【００８５】
次に、専門用語抽出部１３０４が、カタカナ用語リスト１５０３と、全字種用語リスト１５０４とにおいて共起の強い全字種用語を専門用語として特定する専門用語特定処理については、図５と同様である。ここで、シンプソン係数値とＴＦ・ＩＤＦ法とを用いた専門用語特定処理の具体的な例について、図１１に基づき説明する。
【００８６】
図１１は、実施例２に係る全字種用語リスト１５０４から抽出された全字種用語を示す図である。ここで、ロボット工学関連分野のＷｅｂサイト２０のＷｅｂページデータから抽出した全字種用語は、第１キー：シンプソン係数値、第２キー：ＴＦ・ＩＤＦ値でソートしている。そして、シンプソン係数値を求める際に、全字種用語の単独ヒット数の閾値を８に設定し、低頻度な全字種用語を除いてある。又、全字種用語のＴＦ・ＩＤＦ値について、閾値を１５に設定し、どのカタカナ用語とも共起するような全字種用語を除いてある。このようにして、共起ヒット情報に基づいて、「ロボ」、「ゲーム」、「大会」、・・・といった、ロボット工学関連分野における全字種用語を特定している。
【００８７】
そして、専門用語抽出部１３０４が、共起ヒット情報を基に、管理者が設定した閾値以上の共起の強さを持つ全字種用語を、専門用語として抽出する。ここで、専門用語特定処理後の全字種用語の具体的な例について、図１２に基づき説明する。
【００８８】
図１２は、実施例２に係る専門用語特定処理後の全字種用語を示す図である。
【００８９】
図１２に示すように、全字種用語リスト１５０４の全字種用語から、専門用語として、「ロボ、ゲーム、大会、歩行、ＡＳＩＭＯ（登録商標）、ソニー（登録商標）・・・」といった全字種用語が抽出されている。又、全字種用語であったが、単独ヒット数の閾値を設定することで、「アリーナ、ポケモン（登録商標）、ユニーク、・・・」といった全字種用語を、専門用語とすることが回避されている。さらに、ＴＦ・ＩＤＦ値を用いて、「Ｃｏｐｙｒｉｇｈｔ、ＴＯＫＹＯ、ｈｔｔｐ、・・・」といった全字種用語を、専門用語とすることが回避されている。そして、抽出した専門用語と共に、カタカナ用語リスト１５０３の用語を専門用語として、ロボット工学関連分野の専門用語辞書に登録してもよい。さらに、専門用語をロボット工学関連分野の情報を収集するキーワードとして用いるなど、様々なことに用いることができる。
【００９０】
［共起ヒット情報の別の計算方法］
以上、共起ヒット情報の計算方法は、シンプソン係数値とＴＦ・ＩＤＦ法とを用いて説

明したが、シンプソン係数値に代えて、相互情報量値、ダイス係数値、ジャガード係数値、コサイン類似度値を用いてもよい。ここで、カタカナ用語Ｘと全字種用語Ｙの共起の強さを示す、相互情報量値は、数１５のように定義される。

【数１５】

$$R(X,Y)=\log\frac{N|X\cap Y|}{|X||Y|}$$

｜Ｘ｜：カタカナ用語Ｘの単独ヒット数

｜Ｙ｜：全字種用語Ｙの単独ヒット数

｜Ｘ∩Ｙ｜：カタカナ用語Ｘと全字種用語ＹとのＡＮＤ検索でのヒット数

Ｎ：総ページ数

【００９１】

次に、カタカナ用語Ｘと全字種用語Ｙの共起の強さを示す、ダイス係数値は、数１６のように定義される。

【数１６】

$$R(X,Y)=\frac{2|X\cap Y|}{|X|+|Y|}$$

｜Ｘ｜：カタカナ用語Ｘの単独ヒット数

｜Ｙ｜：全字種用語Ｙの単独ヒット数

｜Ｘ∩Ｙ｜：カタカナ用語Ｘと全字種用語ＹとのＡＮＤ検索でのヒット数

【００９２】

次に、カタカナ用語Ｘと全字種用語Ｙの共起の強さを示す、ジャガード係数値は、数１７のように定義される。

【数１７】

$$R(X,Y)=\frac{|X\cap Y|}{|X\cup Y|}$$

｜Ｘ∩Ｙ｜：カタカナ用語Ｘと全字種用語ＹとのＡＮＤ検索でのヒット数

｜Ｘ∪Ｙ｜：カタカナ用語Ｘと全字種用語ＹのＯＲ検索でのヒット数

【００９３】

次に、カタカナ用語Ｘと全字種用語Ｙの共起の強さを示す、コサイン類似度値は、数１８のように定義される。

【数１８】

$$R(X,Y)=\frac{|X\cap Y|}{\sqrt{|X||Y|}}$$

｜Ｘ｜：カタカナ用語Ｘの単独ヒット数
｜Ｙ｜：全字種用語Ｙの単独ヒット数
｜Ｘ∩Ｙ｜：カタカナ用語Ｘと全字種用語ＹとのＡＮＤ検索でのヒット数

【００９４】

［専門用語抽出装置のハードウェア構成］

図１３は、本発明の一実施形態に係る専門用語抽出装置１０（以下、単に専門用語抽出装置と呼ぶ）のハードウェア構成を示す図である。

【００９５】

専門用語抽出装置は、制御部１３０を構成するＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）１３１（マルチプロセッサ構成ではＣＰＵ１３２など複数のＣＰＵが追加されてもよい）、バスライン１０５、通信Ｉ／Ｆ（Ｉ／Ｆ：インターフェイス）１２０、メインメモリ１７０、ＢＩＯＳ（Ｂａｓｉｃ　Ｉｎｐｕｔ　Ｏｕｔｐｕｔ　Ｓｙｓｔｅｍ）１８０、ＵＳＢポート１９０、Ｉ／Ｏコントローラ１６０、キーボード及びマウスなどの入力装置１１０、並びに表示装置１４０を備える。

【００９６】

Ｉ／Ｏコントローラ１６０には、テープドライブ１５１、ハードディスク１５３、光ディスクドライブ１５２、及び半導体メモリ１５４などの記憶部１５０を接続することができる。

【００９７】

ＢＩＯＳ１８０は、専門用語抽出装置の起動時にＣＰＵ１３１が実行するブートプログラムや、専門用語抽出装置のハードウェアに依存するプログラムなどを格納する。

【００９８】

ハードディスク１５３は、専門用語抽出装置として機能するための各種プログラム及び本発明の機能を実行するプログラムを記憶する。

【００９９】

光ディスクドライブ１５２としては、例えば、ＤＶＤ－ＲＯＭドライブ、ＣＤ－ＲＯＭドライブ、ＤＶＤ－ＲＡＭドライブ、ＣＤ－ＲＡＭドライブを使用することができる。この場合は各ドライブに対応した光ディスク１５２１を使用する。光ディスク１５２１から光ディスクドライブ１５２によりプログラム又はデータを読み取り、Ｉ／Ｏコントローラ１６０を介してメインメモリ１７０又はハードディスク１５３に提供することもできる。又、同様にテープドライブ１５１に対応したテープメディア１５１１を主としてバックアップのために使用することもできる。

【０１００】

専門用語抽出装置に提供されるプログラムは、ハードディスク１５３、光ディスク１５２１、又はメモリーカードなどの記録媒体に格納されて提供される。このプログラムは、Ｉ／Ｏコントローラ１６０を介して、記録媒体から読み出され、又は通信Ｉ／Ｆ１２０を介してダウンロードされることによって、専門用語抽出装置にインストールされ実行されてもよい。

【０１０１】

上述のプログラムは、内部又は外部の記憶媒体に格納されてもよい。ここで、記憶媒体としては、ハードディスク１５３、光ディスク１５２１、又はメモリーカードの他に、ＭＤなどの光磁気記録媒体、テープメディア１５１１を用いることができる。又、専用通信回線やインターネットなどの通信回線に接続されたサーバシステムに設けたハードディスク１５３又は光ディスクライブラリなどの記憶装置を記録媒体として使用し、通信ネットワーク３０を介してプログラムを専門用語抽出装置に提供してもよい。

【０１０２】

ここで、表示装置１４０は、ユーザによるデータの入力を受け付ける画面を表示したり、専門用語抽出装置による演算処理結果の画面を表示したりするものであり、ブラウン管表示装置（ＣＲＴ）、液晶表示装置（ＬＣＤ）などのディスプレイ装置を含む。

【０１０３】

ここで、入力装置１１０は、ユーザによる入力の受け付けを行うものであり、キーボード及びマウスなどにより構成してよい。

【０１０４】

又、通信Ｉ／Ｆ１２０は、専門用語抽出装置を専用ネットワーク又は公共ネットワークを介して端末と接続できるようにするためのネットワーク・アダプタである。通信Ｉ／Ｆ１２０は、モデム、ケーブル・モデム及びイーサネット（登録商標）・アダプタを含んでよい。

【０１０５】

以上の例は、専門用語抽出装置のハードウェア構成について主に説明したが、コンピュータに、プログラムをインストールして、そのコンピュータを専門用語抽出装置として動作させることにより上記で説明した機能を実現することもできる。従って、本発明において一実施形態として説明した専門用語抽出装置により実現される機能は、上述の方法を当該コンピュータにより実行することにより、あるいは、上述のプログラムを当該コンピュータに導入して実行することによっても実現可能である。

【０１０６】

以上、本発明の実施形態について説明したが、本発明は上述した実施形態に限るものではない。又、本発明の実施形態に記載された効果は、本発明から生じる最も好適な効果を列挙したに過ぎず、本発明による効果は、本発明の実施例に記載されたものに限定されるものではない。

【図面の簡単な説明】

【０１０７】

【図１】本発明の一実施形態に係るシステム１の全体構成を示す図である。

【図２】本発明の一実施形態に係る専門用語抽出装置１０の機能ブロック図である。

【図３】本発明の一実施形態に係る専門用語抽出処理のフローチャートである。

【図４】本発明の一実施形態に係るカタカナ用語特定処理のフローチャートである。

【図５】本発明の一実施形態に係る専門用語特定処理のフローチャートである。

【図６】実施例１に係るカタカナ語彙を示す図である。

【図７】実施例１に係る共起ヒットの具体例を示す図である。

【図８】実施例１に係る全字種用語リスト１５０４の全字種用語をシンプソン係数値で降順にソートした図である。

【図９】実施例１に係る全字種用語リスト１５０４から抽出された全字種用語を示す図である。

【図１０】実施例１に係る専門用語特定処理後の全字種用語を示す図である。

【図１１】実施例２に係る全字種用語リスト１５０４から抽出された全字種用語を示す図である。

【図１２】実施例２に係る専門用語特定処理後の全字種用語を示す図である。

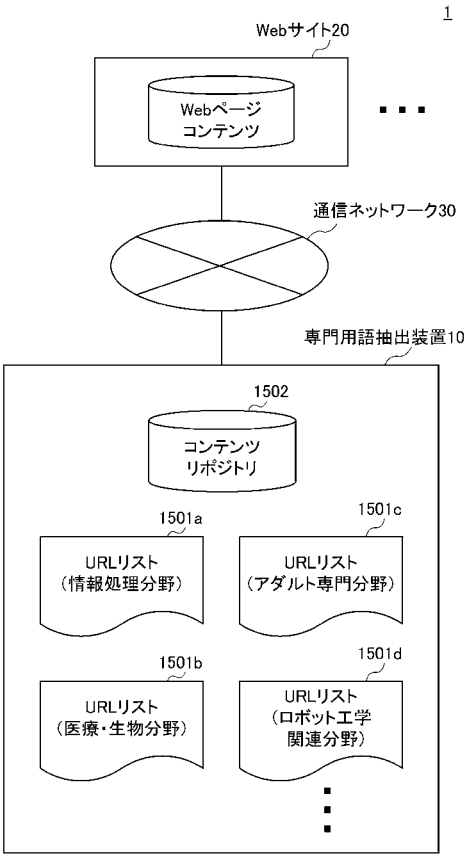
【図１３】本発明の一実施形態に係る専門用語抽出装置１０のハードウェア構成を示す図である。

【符号の説明】

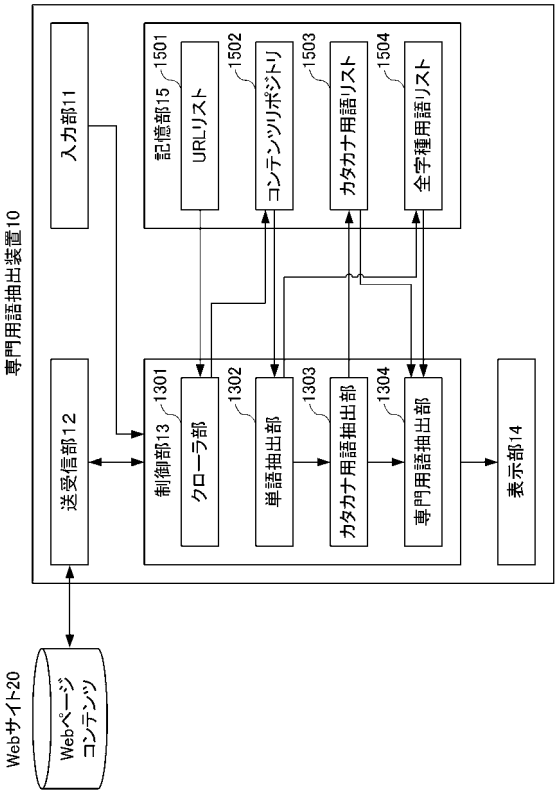
【０１０８】

１　システム
１０　専門用語抽出装置
２０　Ｗｅｂサイト
３０　通信ネットワーク
１５０１　ＵＲＬリスト
１５０２　コンテンツリポジトリ
１５０３　カタカナ用語リスト
１５０４　全字種用語リスト

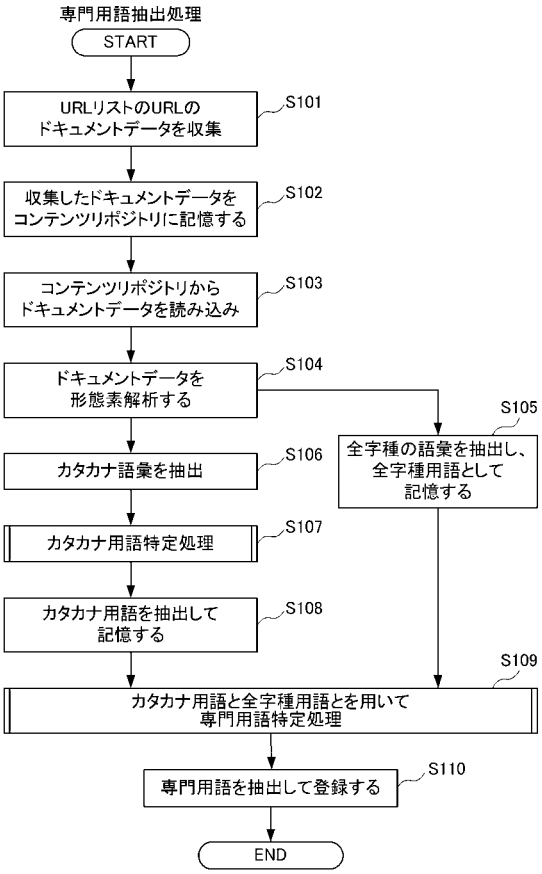
【図１】


1
Webサイト20
Webページ
コンテンツ
通信ネットワーク30
専門用語抽出装置10
1502
コンテンツ
リポジトリ
1501a
URLリスト
（情報処理分野）
1501c
URLリスト
（アダルト専門分野）
1501b
URLリスト
（医療・生物分野）
1501d
URLリスト
（ロボット工学
関連分野）

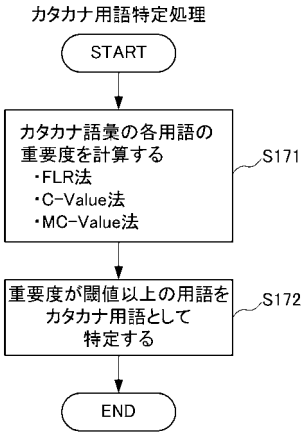
【図２】


専門用語抽出装置10
入力部11
記憶部15
1501
URLリスト
1502
コンテンツリポジトリ
1503
カタカナ用語リスト
1504
全字種用語リスト
送受信部12
制御部13
1301
クローラ部
1302
単語抽出部
1303
カタカナ用語抽出部
1304
専門用語抽出部
表示部14
Webサイト20
Webページ
コンテンツ

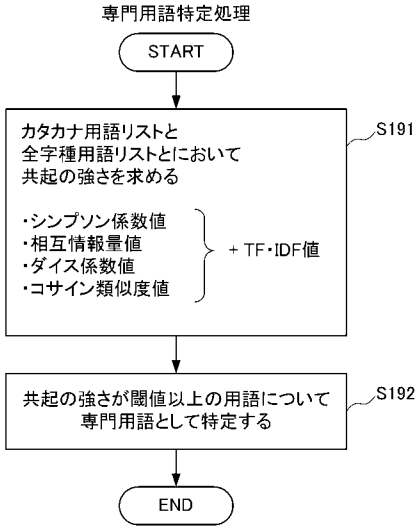
【図３】


専門用語抽出処理
START
URLリストのURLの
ドキュメントデータを収集
S101
収集したドキュメントデータを
コンテンツリポジトリに記憶する
S102
コンテンツリポジトリから
ドキュメントデータを読み込み
S103
ドキュメントデータを
形態素解析する
S104
S105
全字種の語彙を抽出し、
全字種用語として
記憶する
カタカナ語彙を抽出
S106
カタカナ用語特定処理
S107
カタカナ用語を抽出して
記憶する
S108
S109
カタカナ用語と全字種用語とを用いて
専門用語特定処理
専門用語を抽出して登録する
S110
END

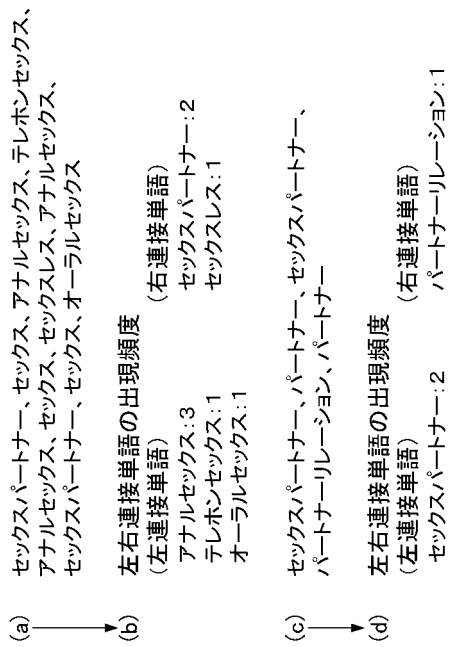
【図４】


カタカナ用語特定処理
START
カタカナ語彙の各用語の
重要度を計算する
・FLR法
・C-Value法
・MC-Value法
S171
重要度が閾値以上の用語を
カタカナ用語として
特定する
S172
END

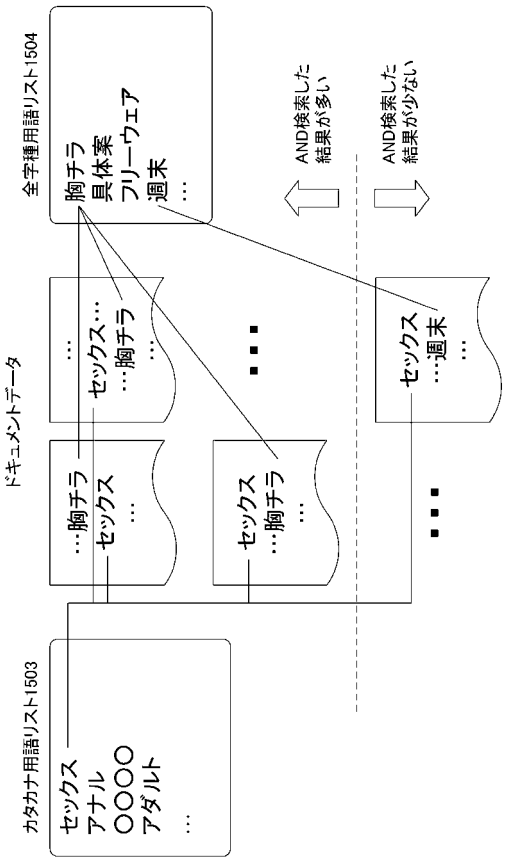
【図５】


専門用語特定処理
START
カタカナ用語リストと
全字種用語リストとにおいて
共起の強さを求める
・シンプソン係数値
・相互情報量値
・ダイス係数値
・コサイン類似度値
＋TF・IDF値
S191
共起の強さが閾値以上の用語について
専門用語として特定する
S192
END

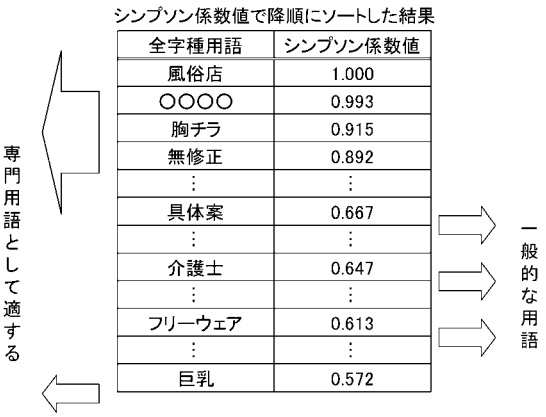
【図６】


(a) セックスパートナー、セックス、アナルセックス、テレホンセックス、
アナルセックス、セックス、セックスレス、アナルセックス、
セックスパートナー、セックス、オーラルセックス
(b) 左右連接単語の出現頻度
(左連接単語) (右連接単語)
アナルセックス:3 セックスパートナー:2
テレホンセックス:1 セックスレス:1
オーラルセックス:1
(c) セックスパートナー、パートナー、セックスパートナー、
パートナーリレーション、パートナー
(d) 左右連接単語の出現頻度
(左連接単語) (右連接単語)
セックスパートナー:2 パートナーリレーション:1

【図７】


ドキュメントデータ
全字種用語リスト1504
胸チラ
具体案
フリーウェア
週末
…
AND検索した結果が多い
AND検索した結果が少ない
…胸チラ
セックス…
…週末
…胸チラ
セックス
…
セックス
…胸チラ
…
セックス
…週末
…
カタカナ用語リスト1503
セックス
アナル
○○○○
アダルト
…

【図８】


シンプソン係数値で降順にソートした結果
全字種用語
シンプソン係数値
風俗店 1.000
○○○○ 0.993
胸チラ 0.915
無修正 0.892
具体案 0.667
介護士 0.647
フリーウェア 0.613
巨乳 0.572
専門用語らしさが高い
一般的な用語

【図９】

URLリスト：アダルト専門分野のWebサイト

全字種用語リストから抽出された全字種用語

| 全字種用語 | シンプソン係数値 | TF・IDF値 |
|---|---|---|
| 風俗店 | 1.000 | 13941.885 |
| 女王 | 1.000 | 10984.916 |
| 風俗嬢 | 1.000 | 9595.345 |
| ○○○○○ | 1.000 | 6233.887 |
| 女性器 | 1.000 | 5084.160 |
| 同性 | 1.000 | 4477.127 |
| 処女膜 | 1.000 | 3296.157 |
| 男性器 | 1.000 | 3198.379 |
| 下着姿 | 1.000 | 2821.487 |
| 性風俗 | 1.000 | 2703.810 |

降順ソート
第1キー：シンプソン係数値
第2キー：TF・IDF値

閾値
単独ヒット数：56
TF・IDF値：760

【図１０】

アダルト専門分野のWebサイトからの専門用語（掲載禁止用語）抽出

| 抽出できた例 |
|---|
| 風俗店、女王、風俗嬢、○○○○○、女性器、同性、処女膜、男性器、下着姿、性風俗、○○○○、胸チラ、無修正、巨乳 |

| 取り除けた例 |
|---|
| ●単独ヒット数の閾値で取り除けた例<br>具体案、介護士、やすみ、ノコギリ、東欧、搭乗客、豚トロ、鈍化、内科医、売却益、麦秋 |
| ●TF・IDF値で取り除けた例<br>フリーウェア、行楽地、株投資、しずく、カワイイ、撫子、茶の間、薄型、中級編、支払、超特価、浄水 |

【図１１】

URLリスト：ロボット工学関連分野のWebサイト（クロールページ数：47）

全字種用語リストから抽出された全字種用語

| 全字種用語 | シンプソン係数値 | TF・IDF値 |
|---|---|---|
| ロボ | 1.000 | 136.264 |
| ゲーム | 1.000 | 129.693 |
| 大会 | 1.000 | 101.530 |
| 歩行 | 1.000 | 100.058 |
| ASIMO | 1.000 | 94.215 |

降順ソート
第1キー：シンプソン係数値
第2キー：TF・IDF値

閾値
単独ヒット数：8
TF・IDF値：15

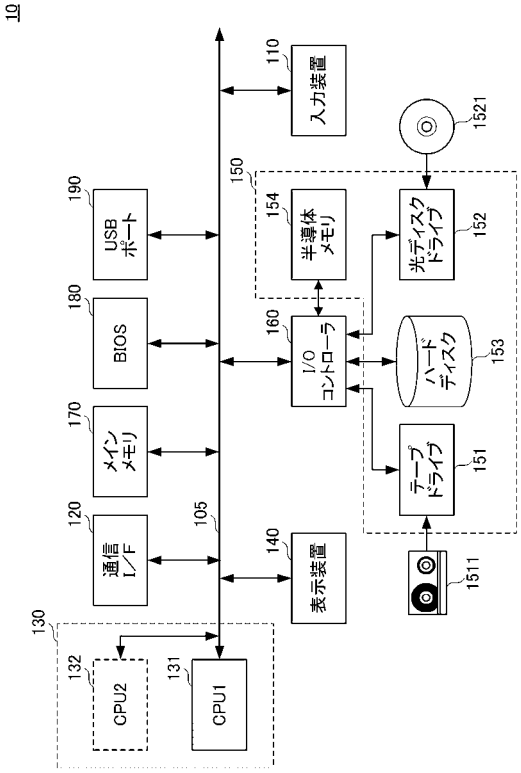
【図１２】

ロボット工学関連分野のWebサイトからの専門用語抽出

| 抽出できた例 |
|---|
| ロボ、ゲーム、大会、歩行、ASIMO、ソニー、設定、デモ、デザイン、センサー、会話、PC、人型 |

| 取り除けた例 |
|---|
| ●単独ヒット数の閾値で取り除けた例<br>アリーナ、ポケモン、ユニーク、意味、完全、玩具、場合、新世界、同様、様子、連載 |
| ●TF・IDF値で取り除けた例<br>Copyright、TOKYO、http、ご紹介、オススメ、フロント、異なる、凶悪犯、超能力、不正、平和 |

【図１３】


10
入力装置
110
1521
光ディスクドライブ
152
半導体メモリ
154
150
USBポート
190
BIOS
180
I/Oコントローラ
160
ハードディスク
153
メインメモリ
170
テープドライブ
151
通信I/F
120
105
表示装置
140
1511
130
CPU2
132
CPU1
131

フロントページの続き

Ｆターム(参考) 5B091 AA15 AB04 AB08 CA02 CC05 CC16